

Der er 50% identitet og 70% similaritet. Der er fire gaps (tre af længde 1 og et af længde 4).

- Da der er tale om aminosyresekvenser, vælger jeg protein BLAST.
- Jeg ved fra foregående spørgsmål, at menneske- og gær-proteinerne er 50% identiske. Derfor er der ikke nogen grund til at bruge PSI-BLAST, så jeg vælger blastp.
- Jeg ved at jeg skal prøve at finde et bestemt Swiss-Prot entry, så jeg sætter databasen til "swissprot".
- Endelig ved jeg at jeg skal lede i organismen bagegær, så jeg indtaster "*Saccharomyces cerevisiae*" under Organism.

E-værdien er **2e-68** (særdeles signifikant). Der er **50%** identitet og **70%** similaritet. Der er **fire** gaps (tre af længde 1 og et af længde 4)—faktisk er alignmentet identisk med det i foregående spørgsmål, bortset fra at allersidste position ikke er med.

Opgave 2 - Karakterisering af KDEL motivet fra dyr og svampe

Spørgsmål a: 43707

Søgestreng: `locations:(location:"endoplasmic reticulum")`

Spørgsmål b: 1538

Søgestreng: `locations:(location:"endoplasmic reticulum"
evidence:experimental)`

NB: i 2014 var der 6642 hits! Mange annoteringer i UniProt er desværre gået fra at være betragtet som "experimental" til at have ukendt evidens. Se http://www.uniprot.org/help/evidences_in_swissprot

Spørgsmål c: 9854

Søgestreng:

`locations:(location:"endoplasmic reticulum lumen")`

eller

`locations:(location:"endoplasmic reticulum")`

`locations:(location:lumen)`

NB: Man kan også få 9748 hits med søgestrengen

`locations:(location:"endoplasmic reticulum lumen" evidence:any)`

hvis man har brugt det grafiske interface til at sætte Evidence tilbage til "Any assertion method". Det skyldes at entries med ukendt evidens ikke kommer med, når der står "evidence:any". Men det er en meget ulogisk opførsel fra interfacets side, og det bør rettes i UniProt!

Bemærk at spørgsmål d, e, f og h også kan give alternative (korrekte) svar ved tilføjelse af "evidence:any".

Spørgsmål d: 5144

Søgestreng: `locations:(location:"endoplasmic reticulum")
annotation:(type:signal)`

Spørgsmål e: 35803

Søgestreng: `locations:(location:"endoplasmic reticulum") fragment:no`

Spørgsmål f: 600

Søgestreng: `locations:(location:"endoplasmic reticulum lumen")
annotation:(type:signal) fragment:no`

Spørgsmål g:

Menneske (*Homo sapiens*) tilhører riget **Metazoa** (dyr).

Bagegær (*Saccharomyces cerevisiae*) tilhører riget **Fungi** (svampe)

Spørgsmål h:

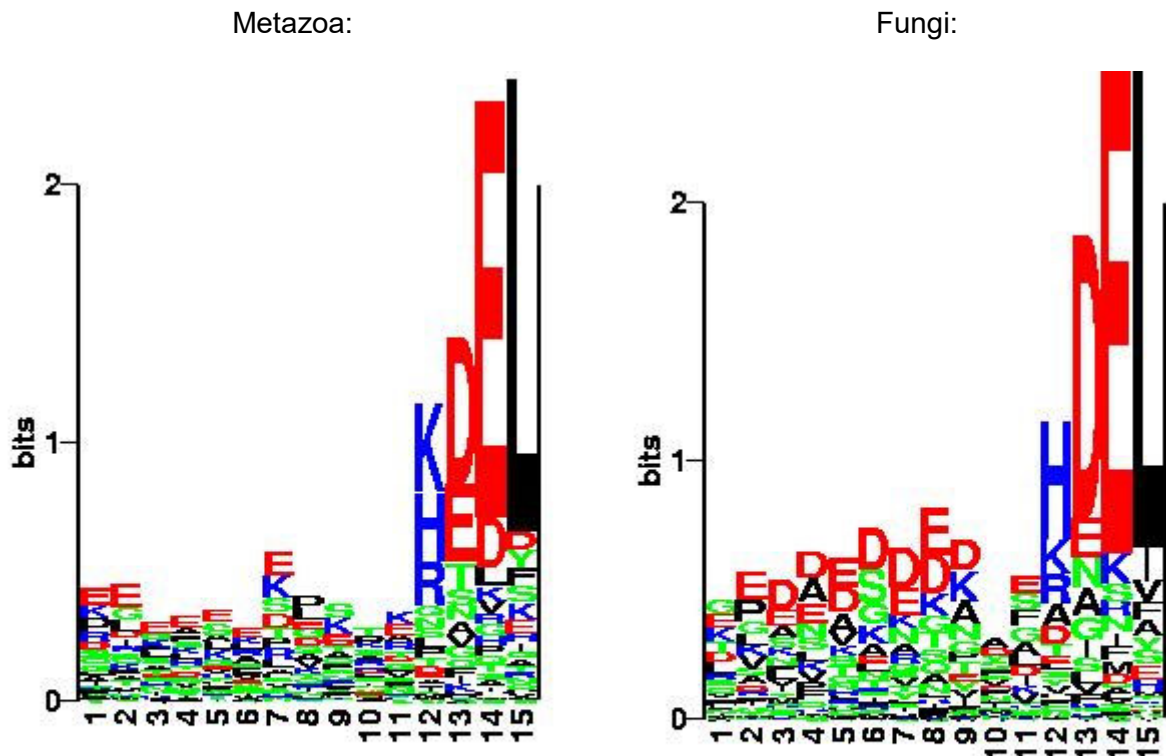
Metazoa: 383

Søgestreng: `locations:(location:"endoplasmic reticulum lumen")`
`annotation:(type:signal) fragment:no taxonomy:metazoa`

Fungi: **57**

Søgestreng: `locations:(location:"endoplasmic reticulum lumen")`
`annotation:(type:signal) fragment:no taxonomy:fungi`

Spørgsmål i:



Positionerne 14 og 15 indeholder mest information i begge logoer. Man kan direkte aflæse konsensus-sekvenserne for KDEL-motiverne ved at tage de øverste bogstaver i de sidste fire positioner (12-15): KDEL for Metazoa og HDEL for Fungi. Forskellen mellem de to riger er tydeligst i position 12, hvor H er meget mere sandsynlig end K for Fungi, men man kan også se en forskel i position 13, D er meget mere sandsynlig end E for Fungi, mens den kun er lidt mere sandsynlig end E for Metazoa. Der er også en forskel i positionerne 1-9, hvor der er en vis præference for D og E i Fungi, men kun en ganske svag præference for E i Metazoa.

Opgave 3 - Fylogeni af KDEL receptoren

Spørgsmål a: Hvis man selv vil downloade sekvenserne i FASTA format fra UniProt, skal man bruge følgende søgestreng:

`name:"ER lumen protein retaining receptor" AND reviewed:yes`

og derefter klikke på "Download" knappen og vælge "FASTA (canonical)". Men det er helt OK blot at bruge den fil, der blev opgivet. *NB: hvis man downloader selv, får man 34 sekvenser i stedet for 33 som i det opgivne sæt — der er kommet én mere siden 2014.*

Efter de beskrevne rettelser ser FASTA filen således ud:

```
>ER21A_XENLA_Metazoa
MNIFRFLGDISHLSAIFILLLLKIWKSRSCAGISGKSQLLFAIVFTARYLDLFTNYISFYN
TSMKVVVYVASSYATVWMIYSKFKATYDGNHDTFRVEFLIVPTAILAFLVNHDFTPLEIFW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGIYRTLYLFNWIWRYQFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGKKLSLPA
>ER21B_XENLA_Metazoa
MNIFRFLGDISHLSAIIILLLLKIWKSRSCAGISGKSQLLFAIVFTTRYLDLFTNFISFYN
TSMKVVVYVASSYATVWMIYSKFKATYDGNHDTFRVEFLIVPTAILSFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGIYRTLYLFNWIWRYQFEEFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGKKLSLPA
>ERD21_BOVIN_Metazoa
MNLFRFLGDLSHLLAIIILLLLKIWKSRSCAGISGKSQVLFVAVFTARYLDLFTNYISLYN
TCMKVVYIACSFTTVWMIYSKFKATYDGNHDTFRVEFLVPTAILAFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGVYRTLYLFNWIWRYHFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGKKLSLPA
>ERD21_HUMAN_Metazoa
MNLFRFLGDLSHLLAIIILLLLKIWKSRSCAGISGKSQVLFVAVFTARYLDLFTNYISLYN
TCMKVVYIACSFTTVWLIYSKFKATYDGNHDTFRVEFLVPTAILAFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGVYRTLYLFNWIWRYHFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGKKLSLPA
>ERD21_MOUSE_Metazoa
MNLFRFLGDLSHLLAIIILLLLKIWKSRSCAGISGKSQVLFVAVFTARYLDLFTNYISLYN
TCMKVVYIACSFTTVWMIYSKFKATYDGNHDTFRVEFLVPTAILAFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGVYRTLYLFNWIWRYHFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGKKLSLPA
>ERD21_RAT_Metazoa
MNLFRFLGDLSHLLAIIILLLLKIWKSRSCAGISGKSQVLFVAVFTARYLDLFTNYISLYN
TCMKVVYIACSFTTVWMIYSKFKATYDGNHDTFRVEFLVPTAVLAFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGVYRTLYLFNWIWRYHFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGKKLSLPA
>ERD21_XENTR_Metazoa
MNIFRFLGDISHLSAILILLLLKIWKSRSCAGISGKSQLLFAIVFTTRYLDLFTNFISLYN
TSMKMVYVASSYATIWMYISKFKATYDGNHDTFRVEFLIVPTAILAFLVNHDFTPLEILW
TFSIYLESVAILPQLFMVSKTGEAETITSHYLFALGIYRALYLFNWIWRYQFEGFFDLIA
IVAGLVQTVLYCDDFFLYITKVLKGKKLSLPA
>ERD22_BOVIN_Metazoa
MNIFRLTGDLSHLAAIVILLLLKIWKTRSCAGISGKSQLLFALVFTRYLDLFTSFISLYN
TSMKLIYIACSYATVYLIYMKFKATYDGNHDTFRVEFLVVPVGGLSFLVNHDFSPLEILW
TFSIYLESVAILPQLFMISKTEAETITTHYLFFLGLYRALYLVNWIWRIFYEGFFDLIA
VVAGVVQTILYCDFFLYITKVLKGKKLSLPA
>ERD22_CHICK_Metazoa
MNIFRLTGDLSHLAAIIILLLLKIWKSRSCAGISGKSQLLFALVFTRYLDLFTSFISLYN
TSMKLIYIACSYATVYLIYMKFKATYDGNHDTFRVEFLVVPVGGLSFLVNHDFSPLEILW
TFSIYLESVAILPQLFMISKTEAETITTHYLFFLGLYRALYLVNWIWRIFYEGFFDLIA
VVAGVVQTILYCDFFLYITKVLKGKKLSLPA
>ERD22_HUMAN_Metazoa
MNIFRLTGDLSHLAAIVILLLLKIWKTRSCAGISGKSQLLFALVFTRYLDLFTSFISLYN
TSMKVIYIACSYATVYLIYKFKATYDGNHDTFRVEFLVVPVGGLSFLVNHDFSPLEILW
TFSIYLESVAILPQLFMISKTEAETITTHYLFFLGLYRALYLVNWIWRIFYEGFFDLIA
VVAGVVQTILYCDFFLYITKVLKGKKLSLPA
>ERD22_DANRE_Metazoa
MNIFRLTGDLSHLAAIIILLLLKIWKSRSCAGISGKSQILFALVFTRYLDLFTSFISLYN
TCMKVIYIGCAYATVYLIYAKFRATYDGNHDTFRAEFLVVPVGGLAFLVNHDFSPLEILW
TFSIYLESVAILPQLFMISKTEAETITTHYLFCGLGYRALYLVNWIWRIFYEGFFDMIA
IVAGVVQTILYCDFFLYITKVLKGKKLSLPA
>ERD22_MOUSE_Metazoa
MNIFRLTGDLSHLAAIVILLLLKIWKTRSCAGISGKSQLLFALVFTRYLDLFTSFISLYN
TSMKLIYIACSYATVYLIYMKFKATYDGNHDTFRVEFLVVPVGGLSFLVNHDFSPLEILW
TFSIYLESVAILPQLFMISKTEAETITTHYLFFLGLYRALYLVNWIWRIFYEGFFDLIA
VVAGVVQTILYCDFFLYITKVLKGKKLSLPA
>ERD22_RAT_Metazoa
MNIFRLTGDLSHLAAIVILLLLKIWKTRSCAGISGKSQLLFALVFTRYLDLFTSFISLYN
TSMKLIYIACSYATVYLIYMKFKATYDGNHDTFRVEFLVVPVGGLSFLVNHDFSPLEILW
TFSIYLESVAILPQLFMISKTEAETITTHYLFFLGLYRALYLVNWIWRIFYEGFFDLIA
VVAGVVQTILYCDFFLYITKVLKGKKLSLPA
```

>ERD22_XENLA_Metazoa
 MNVFRLSGDLCHLAAIIILLKKIWNRSRSCAGISGKSQLLFAMVFTTRYLDLFTSFISLYN
 TSMKVIYMGCAATVYLIYMKFKATYDGNHDTFRVEFLVVPVGGLSVLVNHDFSPLEILW
 TFSIYLESVAIPLQLFMISKTEGAEETITTHYLFFLGLYRALYLFNWIWRFSFEGFFDLIA
 IVAGVVQITILYCDFFLYVTKVLKGKLSLPA

>ERD22_XENTR_Metazoa
 MNVFRLSGDLSHLAAIIILLKKIWNRSRSCAGISGKSQLLFALVFTTRYLDLLTSFISLYN
 TSMKVIYIGCAATVYLIYMKFKATYDGNHDTFRVEFLVVPVGGLSVLVNHDFSPLEILW
 TFSIYLESVAIPLQLFMISKTEGAEETITTHYLFFLGLYRALYLFNWIWRYSFEGFFDLIA
 IVAGVVQITILYCDFFLYVTKVLKGKLSLPA

>ERD23_DANRE_Metazoa
 MNIFRLSGDVCHLAAIIILLFKIWRKSCAGISGKSQVLFALVFTTRYLDLFTSYISAYN
 TVMKVVYLLLAYSTVGLIFFRNSYDSESDSFRVEFLVVPVAGLSFLENYAFTPLEILW
 TFSIYLESVAIPLQLFMITKTGEAESITAHYLLFLGLYRALYLANWLWRFHTEGFDQIA
 VVSGVVQITIFYCDFFLYFTRVLRGSGKMSLPMPV

>ERD23_HUMAN_Metazoa
 MNVFRILGDLSHLLAMILLGKIWRKCKGISGKSQILFALVFTTRYLDLFTNFISIYN
 TVMKVVFLLCAYVTVMYIYKFRKTFDSENDTFRLEFLLVPVIGLSFLENYSFTLLEILW
 TFSIYLESVAIPLQLFMISKTEGAEETITTHYLFFLGLYRALYLANWIRRYQTENFYDQIA
 VVSGVVQITIFYCDFFLYVTKVLKGKLSLPMPI

>ERD23_MOUSE_Metazoa
 MNVFRILGDLSHLLAMILLVKIWRKSCAGISGKSQILFALVFTTRYLDLFSNFISIYN
 TVMKVVFLLCAYVTVMYIYKFRKTFDIENDTFRLEFLLVPVIGLSFLENYSYTPMEVLW
 TFSIYLESVAIPLQLFMISKTEGAEETITTHYLFFLGLYRLLYLANWIRRYQTENFYDQIS
 VVSGVVQITIFYCDFFLYVTKVLKGKLSLPVPV

>ERD23_XENLA_Metazoa
 MNIFRILGDLVSHLLAAIIILLKMWKSKSCAGISGKSQLLFALVFTTRYLDLFTVFISPYN
 TVMKIIFLACAYVTVYLIYKLRKSYDSENDTFRLEFLLVPVIGLSFLENYEFTPLEILW
 TFSIYLESVAIPLQLFMISKTEGAEESITTHYLFFLGLYRVLYLANWIWRHYHTEKFDQIA
 VVSGVVQITIFYDFFFLYVTKVLKGKLSLPMMPV

>ERD23_XENTR_Metazoa
 MNVFRISGDLVSHLLAAIIILLKMWKSKSCAGISGKSQLLFALVFTTRYLDLFTVFISAYN
 TVMKIVFLVCAVTVYLIYKFRKAYDSENDTFRLEFLLVPVIGLSFLENYEFTPLEILW
 TFSIYLESVAIPLQLFMISKTEGAEESITTHYLFFLGLYRVLYLANWIWRHYHTEKFDQIA
 VVSGVVQITIFYDFFFLYITKVLKGKLSLPMMPV

>ERD2_ARATH_Viridiplantae
 MNIFRFAGDMSHLISVLILLKKIYATKSCAGISLKTQELYALVFLTRYLDLFTDYVSLYN
 SIMKIVFIASSLAIVCMRRHPLVRRSYDKDLDTFRHQYVVLACFVLGLILNEKFTVQEV
 FWAFSIYLEAVALPQLVLLQSRGNVDNLGTQYVFLGAYRGLYIINWIYRYFTEDHFTTR
 WIACVSGVLQATALYADFFYYYYISWKTNTKLKLP

>ERD2_CAEBR_Metazoa
 MNIFRITADLAHAIVAILLLKKIWKSRSCAGISGRSQILFAVTFTRYLDLFTSFYSLYN
 TVMKVFLFAGSIGTVYLMVVKFKATYDRNNDTFRIEFLVIPSIIILALINHEFMFMEVMW
 TFSIYLEAVALIMPQLFMLSRTGNAETITAHYLFALGSYRFLYIFNWVYRYTESFFDPIA
 VVAGIVQTVLYADFFLYITRVIQSNRQFEMSA

>ERD2_DICDI_Amoebzoa
 MNLFSFLGDMHLHLSMLILLFKIKNDKSCAGVSLKSQILFTIVFTARYLDLFTNYVSLYI
 TFMKITIYIAVSYYTLHLIARKYKFTYDKDHDTFKIVYLIASCAILSLITYDKTTIGIYST
 FLEILWTFISIYLESIAIPLQLILLQRTGEVEALTSNYIVLLGGYRAFYLFNWIYRITFYN
 WSGKIEMLSGLLQITILYADFFYYYAKSRMYGKKLVLPQ

>ERD2_CAEL_Metazoa
 MNLFRFTADVAHAIAIVVLLKKIWKSRSCAGISGRSQLLFALVFVTRYLDLFTNFFSFYN
 TAMKIFYLVASFQTVYLMWAKFKATYDRNNDTFRIEFLVIPSIMILALLINHEFIFMEVMW
 TFSIYLEAVALIMPQLFMLSRTGNAETITAHYLFALGSYRFLYIILNWVYRYTESFFDPIS
 VVAGIVQTVLYADFFLYITRVIQSNRQFEMSA

>ERD2_DROME_Metazoa
 MNIFRFAGDLSHVFAIIILLKKIWKTRSCAGISGKSQILFAVVYLVTRYLDLFTTYVSLYN
 SVMKVFLFATSGATVYLMYVKFKATYDHNHDSFRIEFLVPCALLSLVINHEFTVMEVLW
 TFSIYLESVAIPLQLFLVSRTEGAEESITSHYLFALGSYRALYLLNWVYRYMVESHYDLIA
 IFAGVVQTVLYCDFFLYITKVLKGKQLLPA

>ERD2_ENTHI_Amoebzoa
 MVFNLFRIADLVHLLSIYFLLTKIISHKNCIGISLRSQILFFIVVWTRYLDIFYNFYSL
 YNTILKIVYLTTSAITYIYLSKRFRATYDKIHDTLNVWYLVPCIVLAFIFTEDYSITEI
 CWTFSIFLEAVALPQILLRSTGEVENLNSQYIFCLGLYRALYIINWIYRYATEQSYWS
 PLTWICGSIQTLLYVEYFYIYIKSRVEGTFKVLVLPY

>ERD2_KLULA_Fungi
 MLNVFRIAGDFSHLASIIILIQSITTSNSVDGISLKTQLLYTLVFITRYLNLFTKWTSLY

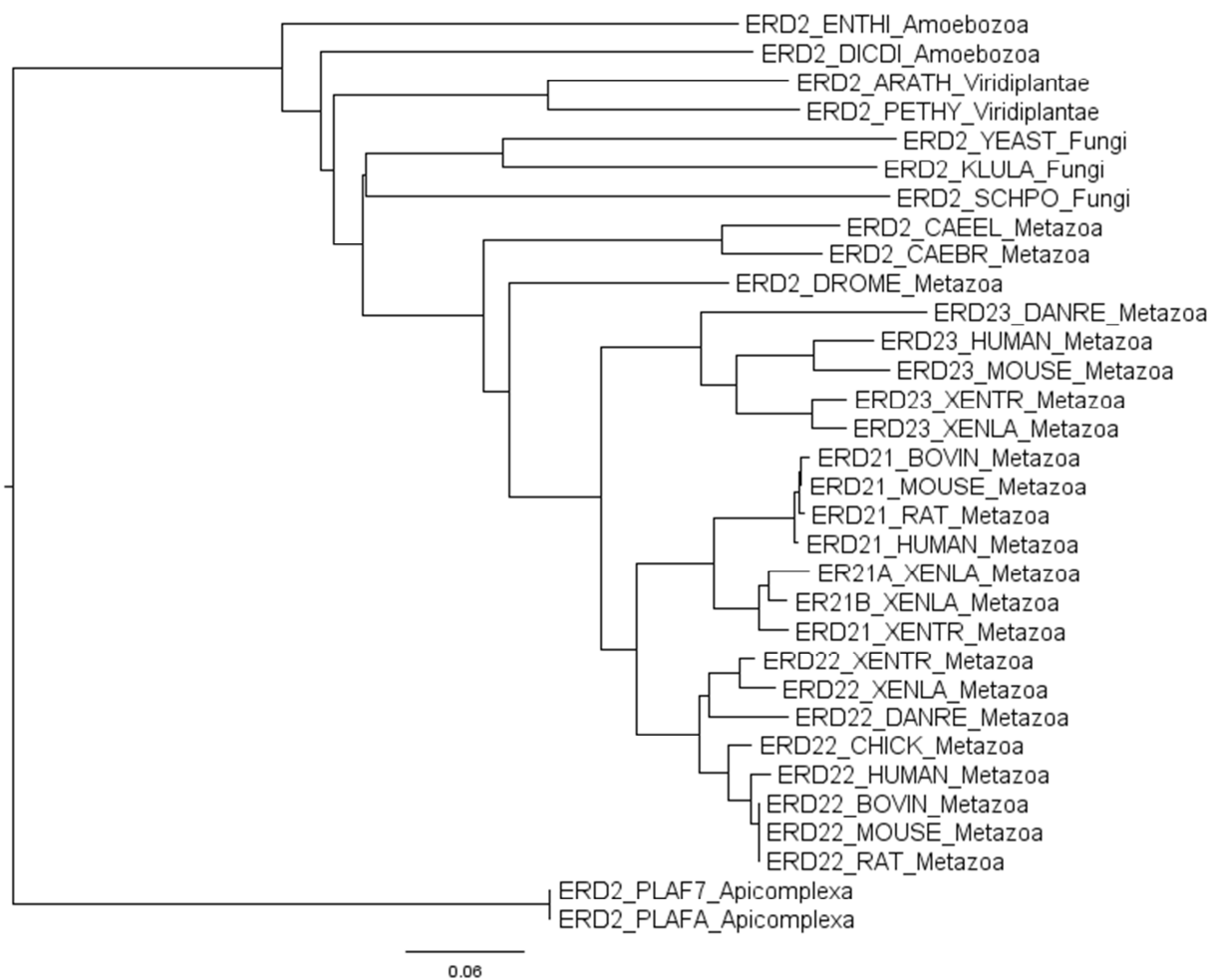
```

NFLMKIVFISSSVYVIVLMRQQKFNVPAYQDMITRDQFKIKFLIVPCILLGLIFNYRFS
FIQICWSFSLWLESVAIILPQLFMLTKTGKAKQLTSHYIFALGLYRALYIPNWIWRYYTEE
RFDKLSVFTGVIQTLVYSDFFYIYYQKVIKLGGDLELPQ
>ERD2_PETHY_Viridiplantae
MNIERLAGDMTHLASVVLVLLKIHTIKSCAGVSLKTQELYALVFVTRYLDIFTDFISLYN
TTMKLVFLGSSLSIVWYMRHKKIVRRSYDKDQDTFRHLFLVLPCLLLALVINEKFTFKEV
MWTFSIYLEAVAILPQLVLLQRTNRIDNLTGQYIFLLGAYRSFYILNWVYRYFTEPHFVH
WITWIAGLIQTLLYADFFYYFQSWKNNTKLELPA
>ERD2_PLAF7_Apicomplexa
MNIERLIGDILHLVSMYILIMKLKSKNCIGISCRMQELYLIVFLCRYIDLFFVFSFYN
TVMKITFILTIAITYILIRLKLPISTYNRKVDNFKSEKYLIPCLVLSLLTCKTYNLYN
ILWSFSIWLESVAIILPQLVLLLEKQREVENITSHYVITMGLYRAFYLINWIYRYFFDDKPY
INVVGWIGGLIQTLLYIDFFYYFALAKWYGKKLVLPFNGEV
>ERD2_PLAFA_Apicomplexa
MNIERLIGDILHLVSMYILIMKLKSKNCIGISCRMQELYLIVFLCRYIDLFFVFSFYN
TVMKITFILTIAITYILIRLKLPISTYNRKVDNFKSEKYLIPCLVLSLLTCKTYNLYN
ILWSFSIWLESVAIILPQLVLLLEKQREVENITSHYVITMGLYRAFYLINWIYRYFFDDKPY
INVVGWIGGLIQTLLYIDFFYYFALAKWYGKKLVLPFNGEV
>ERD2_SCHPO_Fungi
MTFFSALGDMAHLAIFLLHMRKSKSKTCSGLSLKSHLLFLLVYVTRYLNLFWRYKSLYY
FLMRIVFIASESYICYLMLMTLRPTYDKRLDTFRTEYILGGCAVLALIYPTSYTISNILW
TFSIWLESVAIILPQLFMLQRSGETESLTAHYLFAMCLYRGLYIPHWIYRIAVHKKVIGVA
ILAGIIQTVLYGDFAVVYRRTVLQGGKFRLEPA
>ERD2_YEAST_Fungi
MNPFRILGDLSHLTSILILIHNIKTRYIEGISFKTQTLTYALVFITRYLDLLTFHWVSLY
NALMKIFFIVSTAYIVVLLQGSKRNTIAYNEMLMHDTFKIQHLLIGSALMSVFFHHKFT
FLELAWSFSVWLESVAIILPQLYMLSKGGKTRSLTVHYIFAMGLYRALYIPNWIWRYSTED
KKLDKIAFFAGLLQTLLYSDFFYIYYTKVIRKGFKLKP

```

Spørgsmål b: Jeg gik til MAFFT serveren på EBI og lavede et multiple alignment af den rettede FASTA fil. Som output format valgte jeg FASTA. Derefter downloadede jeg resultatet ved at højreklikke på “Download Alignment File” i fanen “Alignments”. Dette resultat uploadede jeg til TreeHugger og klikkede på “Submit query”. Så højreklikkede jeg på “Download data in Newick/Phylip format” og gemte filen på min computer. Så startede jeg FigTree og åbnede den gemte fil.

Jeg blev bedt om at bruge slægten *Plasmodium* som *outgroup*. Der er to *Plasmodium* sekvenser i mit datasæt, ERD2_PLAFA og ERD2_PLAF7. Jeg klikker på den gren, der fører til disse to og klikker på “Reroot” øverst i vinduet. Endelig klikker jeg på “Tip Labels” til venstre i vinduet og sætter “Font Size” til 15. Så ser mit træ således ud (efter at det er blevet eksporteret som .png fil):



Spørgsmål c: Dyr (24 sekvenser), planter (2 sekvenser) og svampe (3 sekvenser) er hver især repræsenteret som systematiske grupper.

Der er én fejl: ERD2_ENTHI (*Entamoeba histolytica*) og ERD2_DICDI (*Dictyostelium discoideum*) burde høre sammen i en systematisk gruppe (*Amoebozoa*), men det gør de ikke.

NB: i 2014 var der en fejl mere, idet svampe ikke var repræsenteret som en systematisk gruppe. Det eneste, der er sket siden dengang, er at MAFFT er blevet opdateret til version 7.215. Det viser hvor meget alignment-metoden kan betyde for den fylogenetiske analyse!

Spørgsmål d: Mulighed **2** er den rigtige: Opdelingen i ER lumen protein retaining receptor 1, 2 og 3 er sket tidligt i hvirveldyrenes (*Vertebrata*) udvikling, før de forskellige klasser (såsom benfisk, padder og pattedyr) spaltede ud. Det kan man se af at ERD21-, ERD22- og ERD23-sekvenserne falder i tre adskilte grupper, der hver især indeholder flere klasser af hvirveldyr. Alle tre grupper indeholder både pattedyr (f.eks. HUMAN) og padder (*Xenopus*), mens to af grupperne (ERD22 og ERD23) også indeholder fisk (*Danio rerio*).

Opgave 4 - Forudsigelse af ER lumen proteiner

Spørgsmål a: Det positive testsæt indeholder 107 linjer, og de første 10 linjer ser således ud:

```
DMEEDDDQKAVKDEL 1
DMEEDDDQKAVKDEL 1
DMEEDDDQKAVKDEL 1
DMEEDDDQKAVKDEL 1
DMEEDDDQKAVKDEL 1
DMEEDDDQKAVKDEL 1
DLEEDDDQKAVRDEL 1
DMEEDDDQKAVKDEL 1
PEPPANSTMGSKEEL 1
PEAQANSTLGPKEEL 1
PEPPANSTMGSKEEL 1
```

Spørgsmål b: I position 13 (den fremhævede position) er der syv D'er, to N'er og ét G. Det giver:

$$f_D = 0.7$$

$$f_N = 0.2$$

$$f_G = 0.1$$

$$g_D = f_D * q(D|D) + f_N * q(D|N) + f_G * q(D|G) = 0.7 * 0.40 + 0.2 * 0.08 + 0.1 * 0.03 = 0.299$$

$$g_E = f_D * q(E|D) + f_N * q(E|N) + f_G * q(E|G) = 0.7 * 0.09 + 0.2 * 0.05 + 0.1 * 0.03 = 0.076$$

$$p_D = (\alpha * f_D + \beta * g_D) / (\alpha + \beta) = (9 * 0.7 + 10 * 0.299) / (9 + 10) = 0.489$$

$$p_E = (\alpha * f_E + \beta * g_E) / (\alpha + \beta) = (9 * 0 + 10 * 0.076) / (9 + 10) = 0.040$$

$$w_D = 2 * \log(p_D / q_D) / \log(2) = 2 * \log(0.489 / 0.054) / \log(2) = 6.358$$

$$w_E = 2 * \log(p_E / q_E) / \log(2) = 2 * \log(0.040 / 0.054) / \log(2) = -0.866$$

Hvis man bruger EasyPred til at kontrollere w-værdierne, får man:

$$w_D = 6.355$$

$$w_E = -0.850$$

Hvilket er tæt nok på, givet at værdierne i tabellen i handoutet er afrundede.

Spørgsmål c: 539786 sekvenser

Søgestreng: `taxonomy:metazoa annotation:(type:signal) fragment:no NOT locations:(location:"endoplasmic reticulum")`

Spørgsmål d: 333 linjer (107 positive og 226 negative)

NB: det negative testsæt er lavet i 2014, da svaret på spørgsmål c var 22649 sekvenser

Spørgsmål e:

Pearson coefficient: 0.64637

Aroc value: 0.85163

Hvis vores metode var perfekt, ville begge værdier være 1. Hvis vores metode gættede tilfældigt, ville Pearson være 0 og Aroc være 0.5.